УДК 8'32, 811.13, 519.765

КВАНТИТАТИВНЫЙ АНАЛИЗ ЛЕКСИКИ ВЕРХНЕ-ЛУЖИЦКОГО ЯЗЫКА

Г. Д. Селезнев, А. А. Кретов

Воронежский государственный университет

Поступила в редакцию 1.03.2009 г.

Аннотация. Исследуются статистические закономерности лексики верхне-лужицкого языка: распределения слов по длине и по числу значений, зависимость числа синонимических рядов от количества слов в ряду, зависимость числа слов от количества образуемых ими фразеосочетаний. Производится аппроксимация полученных зависимостей известными законами распределений.

Ключевые слова: квантитативная лексикология, верхнелужицкий язык, аппроксимация.

Abstract. The article explores statistical regularities in vocabulary of Upper Lusatian language: distribution of words by their length and number of meanings; relation of quantity of synonymic chains to the number of words in a chain; dependence of the number of words on the number of idioms which they form. The approximation of the dependences based on the partition laws is carried out.

Keywords: quantitative lexicology, Upper Lusatian language, approximation.

ВВЕДЕНИЕ

Последовательность этапов любого научного исследования состоит в систематическом сжатии, компактификации информации об изучаемых объектах. Что является существенным, что следует отфильтровать, а что отбросить, как лишнее, зависит от целей исследования. Из огромного исходного материала наблюдений путем их систематизации, статистической обработки, выявления связей и зависимостей иногда удается получить компактную таблицу, а лучше короткую математическую формулу с двумя-тремя числовыми параметрами. Эта формула впитывает в себя самое важное знание об объекте. И тогда говорят – мы открыли Закон. Дальше можно размышлять, почему формула, описывающая исследуемый объект, именно такова и коэффициенты в ней именно такие, сравнивать, строить модели, стараться свести полученный закон к уже известным первым принципам. Так создаются теории.

Примером такого исследования является квантитативный анализ лексики национальных языков. Статистические исследования толковых, частотных или двуязычных словарей позволяют получить существенные сведения об исследуемом языке. О динамике его развития, типологической соотнесенности с другими языками, национальной специфике. Результаты исследований представляются в виде таблиц зависимостей. Как правило, полученные зависимости удается аппроксимировать известными законами распределений с вполне приемлемой точностью.

В предлагаемой работе исследовались: распределения слов по длине и по числу значений слова, зависимость числа синонимических рядов от количества слов в ряду, зависимость числа слов от количества образуемых ими фразеосочетаний. Исследование проведено на материале словаря верхне-лужицкого языка [1]. Это один из языков славянской группы, распространенный среди сербов, проживающих в восточных землях Германии. Исследование является частью работы по исследованию лексикологии славянских и романских языков, которая проводится на кафедре теоретической и прикладной лингвистики воронежского госуниверситета.

1. РАСПРЕДЕЛЕНИЕ ДЛИН СЛОВ

В таблице 1 представлено распределение слов по длине, измеряемой в количестве букв в слове.

Производилась аппроксимация экспериментального распределения Гамма-распределением вида

$$N(l) = \mathbf{N}\frac{\beta^{\alpha}}{\Gamma(\alpha)} l^{\alpha-1}\exp(-\beta l),$$

где $N(l)$ – количество слов длины l, **N** – общее количество слов – объем словаря, α, β – подбираемые, существенно важные параметры распределения, Γ (α) – гамма-функция. При **N** =31933, α = 12. 437, β = 0.672 показатель качества аппроксимации $R^2 = 0.994$. Распределение вероятностей длин слов, как результат аппроксимации представлен на рис. 1.

2. РАСПРЕДЕЛЕНИЕ СЛОВ ПО ЧИСЛУ ЗНАЧЕНИЙ

В таблице 2 представлено распределение слов по числу значений.

Производилась аппроксимация экспериментального распределения экспоненциальным распределением и распределением Ципфа-Мандельброта [2,3].

1) Для экспоненциального распределения вида

$$L(k) = L_0\exp(-\gamma k),$$

где $L(k)$ – количество слов, имеющих k значений, L_0, γ – подбираемые множитель и показатель экспоненты. При L_0, = 75544 и $\gamma \approx 1.467$ показатель качества аппроксимации $R^2 = 0.99$.

2) Для распределения Ципфа-Мандельброта вида

$$L(k) = A(k+B)^{-\lambda},$$

где $L(k)$ – количество слов, имеющих k значений, A, B, λ – подбираемые коэффициенты. При $A = 1.17\cdot10^{21}$, B = 8.063 и $\lambda \approx 17.39$ показатель качества аппроксимации $R^2 = 0.9999$. Результаты аппроксимации представлены на рис. 2 и 3.

Таблица 1.

Распределение слов по длине

Длина слова	Количество слов	Длина слова	Количество слов
1	5	12	1441
2	34	13	770
3	333	14	359
4	1013	15	181
5	2458	16	72
6	3766	17	23
7	4908	18	13
8	5514	19	5
9	4852	20	4
10	3760	21	1
11	2421	22	0
Сумма 31933			

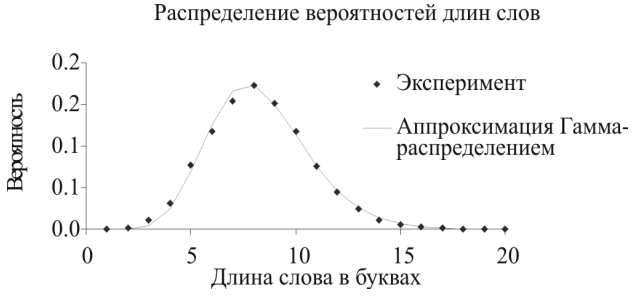

Рис. 1.

Таблица 2.

Распределение слов по числу значений

Количество значений слова	Количество слов
1	26578
2	4302
3	858
4	139
5	42
6	7
7	3
8	1
Сумма 31930	

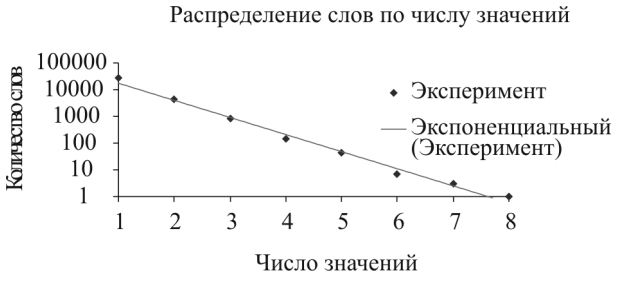

Рис 2.

Таблица 3.

Распределение синонимических рядов.

Слов в ряду	Синонимических рядов
1	32556
2	2415
3	285
4	44
5	6
Сумма 35306	

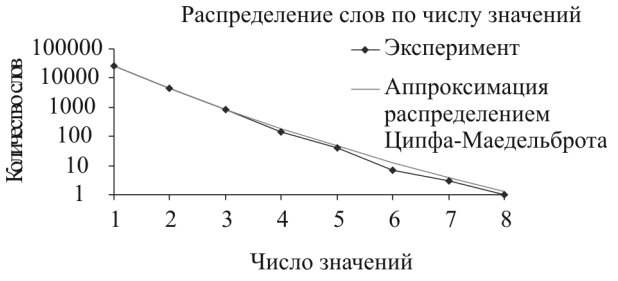

Рис. 3.

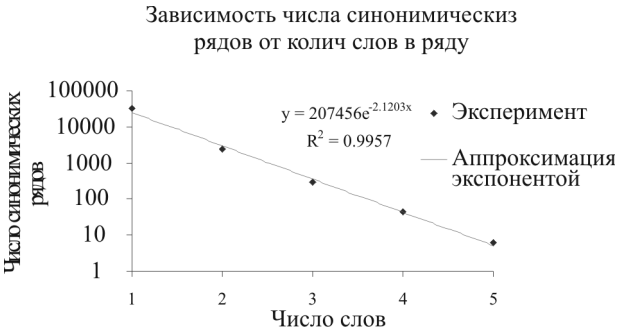

Рис 4.

В пределах ошибки аппроксимации трудно отдать предпочтение какому-то из этих двух распределений.

3. ЗАВИСИМОСТЬ ЧИСЛА СИНОНИМИЧЕСКИХ РЯДОВ ОТ КОЛИЧЕСТВА СЛОВ В РЯДУ

В таблице 3 представлено распределение количества синонимических рядов от количества слов в ряду.

Производилась аппроксимация этого распределения экспоненциальным распределением вида

$$S(m) = S_0 \exp(-\mu m),$$

где $S(m)$ – количество синонимических рядов, имеющих m слов в ряду, S_0, μ – подбираемые множитель и показатель экспоненты. Результат аппроксимации представлен на рис. 2. При $S_0 = 207456$ и $\mu \approx 2.12$ показатель качества аппроксимации $R^2 = 0.996$. Результаты аппроксимации представлены на рисунке 4.

4. ЗАВИСИМОСТЬ ЧИСЛА СЛОВ ОТ КОЛИЧЕСТВА ФРАЗЕОСОЧЕТАНИЙ

В таблице 4 представлено распределение слов по количеству образуемых ими фразеосочетаний. Аппроксимация распределением Ципфа-Мандельброта.

Таблица 4.

Распределение слов по количеству фразеосочетаний.

Фразеосочетаний	Слов	Фразеосочетаний	Слов	Фразеосочетаний	Слов
0	26960	9	16	18	2
1	3146	10	9	19	2
2	924	11	10	22	3
3	403	12	7	23	1
4	204	13	6	25	2
5	103	14	3	28	1
6	66	15	1	34	2
7	34	16	1	36	2
8	21	17	1	41	1
				66	1
Сумма 31932					

Для распределения Ципфа-Мандельброта вида

$$L(k) = A(k + B)^{-\lambda},$$

где $L(k)$ – количество слов, образующих k фразеосочетаний, A, B, λ – подбираемые коэффициенты. При $A = 26960$, $B = 1$ и $\lambda = 3.1$ показатель качества аппроксимации $R^2 = 0.9999$. Результаты аппроксимации представлены на рисунке 5.

ОБСУЖДЕНИЕ РЕЗУЛЬТАТОВ

Из приведенных результатов следует, что исследуемые в работе зависимости подчиняются строгим статистическим законам. Достигаемая точность аппроксимации, во всех случаях не меньшая 99% , позволяет считать, что найденные теоретические распределения действительно адекватны экспериментальным. Численные значения параметров распределений близки к значениям, которые получены при исследованиях словарей других славянских и романских языков [2, 3]. Результат, полученный в разделе 2, подтверждает сделанный в работе [2] вывод о том, что распределение Ципфа-Мандельброта является «промежуточным», «преходным» от экспоненциального к гиперболическому распределению Ципфа.

Рис.5.

Работа выполнена при финансовой поддержке Российского гуманитарного научного фонда (РГНФ) проект 07-04-00036а.

ЛИТЕРАТУРА

1. *Трофимович К. К.* Верхне-лужицко русский словарь 36000 слов Ред. Ф.Михалк и П.Фёлькель. Изд. „Русский язык" Москва, Народное изд. „Домовина" Бауцен 1974.

2. *Селезнев Г.Д.* Природа экспоненциального распределения слов по числу значений. Вестник Воронежского государственного университета. Сер. Лингвистика и межкультурная коммуникация. № 2. 2007. —Воронеж, С. 42-46.

3. *Селезнев Г.Д.* Как распределяются слова по числу значений. Проблемы компьютерной лингвистики: сб. науч. тр. — Воронеж, 2008 .— Вып. 3. - С. 220-225.

СелезневГеннадий Данилович - доц. кафедры теоретической и прикладной лингвистики и кафедры онтологии и теории познания, Воронежский государственный университет.Тел. (4732) 204-149. E-mail: selforg@newmail.ru

Кретов Алексей Александрович – д.фил.н., профессор, зав. каф. структурной и прикладной лингвистики факультета РГФ, Воронежский государственный университет. Тел. (4732)204-149. E-mail: a_a_kretov@rambler.ru

Seleznev Gennady D. - Associate Professor, the dept. of the Theoretical and Applied Lingvistics , Voronezh State University. Tel.(4732) 204-149, E-mail: selforg@newmail.ru

Kretov Aleksey A. - Doctor of Philology Sciences, Professor, the Head of the dept. of Theoretical and Applied Lingvistics , Voronezh State University. Tel.(4732) 204-149. E-mail: a_a_kretov@rambler.ru